Material Teórico - Módulo Noções Básicas de Estatística

Introdução à Estatística

Sétimo Ano do Ensino Fundamental

Autor: Prof. Fabrício Siqueira Benevides
Revisor: Prof. Antonio Caminha M. Neto

20 de julho de 2023

1 Introdução à Estatística

A "Estatística" estuda a coleta, organização, análise, interpretação e apresentação de dados. Algo de extrema importância na sociedade moderna. Ao usar estatísticas, cientistas, pesquisadores e tomadores de decisão podem identificar padrões, validar teorias e fazer escolhas mais bem informados em uma ampla variedade de campos, incluindo Economia, Medicina, Ciências Sociais, estudos ambientais e muito mais.

Em nossa vida cotidiana, encontramos estatísticas o tempo todo. Por exemplo, quando você ouve que "75% das crianças na sua escola gostam de pizza", isso é uma estatística. Quando as previsões meteorológicas prevêem uma chance de 30% de chuva, isso é baseado em modelos estatísticos usando dados meteorológicos anteriores.

Há duas vertentes principais na Estatística:

- A **Estatística Descritiva**, que tem o objetivo de *resumir* grandes conjuntos de dados.
- A **Estatística Inferencial**, que tem o objetivo de *obter conclusões* a partir de dados sujeitos a variações aleatórias.

A **Estatística Descritiva** nos ajuda a entender as propriedades básicas de um conjunto de dados e nos fornece uma visão geral rápida sobre eles, apresentando as informações por meio de tabelas, gráficos, medidas centrais (como a média) e frequências dos dados.

Por exemplo, podemos partir de uma lista com as alturas de cada um dos estudantes de uma universidade que possui cerca de 20 mil alunos. Para simplificar essa informação, podemos calcular a *média* da alturas (somando todas as alturas e dividindo o resultado pela quantidade de alunos).

A **Estatística Inferencial** vai além da descrição de dados e os utiliza para tirar conclusões ou fazer previsões sobre uma população maior. Ele permite que os pesquisadores façam inferências com base em uma amostra de dados, ajudando-os a fazer generalizações sobre todo um grupo ou população. Por exemplo, se não tivermos acesso à lista com todas as

alturas dos alunos, mas conseguirmos obter parte dessa lista por meio de uma pesquisa na qual foram entrevistados apenas alguns dos alunos, e quisermos obter alguma conclusão sobre as alturas dos demais alunos da escola.

Estudar Estatística é importante para que, além da tarefa óbvia de entender o conjunto de dados analisados, possamos também entender os limites da própria Estatística. Por exemplo, ao resumir um conjunto de dados certamente perdemos alguma informação. Saber que a média das alturas do grupo de alunos é 1,65 m nos diz menos do que saber a altura de cada um deles, individualmente. Observado apenas a média, já não conhecemos as alturas dos alunos mais baixos ou mais altos. Sequer sabemos se a maioria dos alunos está próximo da média ou se há muitos alunos mais baixos e outros tantos mais altos. Ainda assim, conhecer a média de um conjunto de valores é uma informação preciosa. Dessa forma, um ponto importante é saber *interpretar* os dados resumidos para tirar conclusões a partir deles (já que também seria impraticável analisar cada uma das 20 mil alturas individualmente).

Da mesma maneira, na Estatística Inferencial, ao tirar conclusões sobre um conjunto de dados maior que a amostra obtida, é importante saber quais são os limites dessas conclusões. Por exemplo, se a altura média dos estudantes da universidade for 1,65 m, podemos inferir que a altura média dos estudantes do país é 1,65 m? A resposta é não. A Estatística Inferencial **não** permite fazer este tipo de generalização.

Por outro lado, ela nos permite fazer pesquisas durante uma eleição, que podem prever com bastante confiança quem será o vencedor entrevistando apenas uma pequena parte dos eleitores. Para isso, é preciso ser bastante criterioso ao selecionar a **amostra** (o conjunto de pessoas que serão entrevistadas), para que ela represente bem a população como um todo (e, para isso, precisamos ter um **censo** confiável, uma base de dados com características gerais de todos os eleitores). Se você fizer uma pesquisa apenas entre seus amigos, quando a eleição tiver escopo nacional, muito provavelmente você falhará em prever o resultado. Por isso, o estudo da Estatística

vai muito além de simplesmente trabalhar com os números. É preciso usar boas técnicas de coleta de dados e de análise para que as conclusões sejam válidas.

Nesta aula, apresentamos alguns dos conceitos básicos estudados em Estatística: como são classificados os dados, as noções de frequência absoluta e relativa e as noções de média, moda e mediana.

Mas lembre-se: o uso de estatísticas não se trata apenas de números; trata-se também de fazer as perguntas certas e ser curioso. Você pode usar estatísticas para aprender, descobrir coisas novas e até se divertir!

2 Tipos de dados

Em estatística, uma variável representa uma característica ou quantidade que pode ser medida, observada ou controlada em um estudo ou experimento. Por exemplo, a altura de cada estudante é uma variável da população de estudantes. Outros exemplos são: a idade, a intenção de voto (numa certa eleição), o local de moradia e a quantidade de pessoas infectadas por certo vírus em um dado momento.

A distinção entre os tipos de variáveis é importante pois diferentes técnicas de análise e de visualização são aplicadas a cada tipo. E a interpretação dos resultados pode variar dependendo da natureza da variável em estudo. As variáveis podem ser classificadas em dois tipos: qualitativas e quantitativas.

1. uma **variável qualitativa** (ou categórica) descreve uma característica com base em categorias que não precisam possuir uma ordem específica. Exemplos incluem o gênero, a cor dos olhos, o estado civil e a classe social.

2. uma **variável quantitativa** (ou numérica) representa uma quantidade numérica e elas podem ser subdivididas em 2 tipos: contínuas e discretas.

   (a) *variáveis contínuas* podem assumir qualquer valor numérico dentro de um dado intervalo de números reais. Por exemplo: altura, peso e temperatura.

(b) *variáveis discretas* podem assumir apenas valores distintos e separados. Normalmente (mas nem sempre) ela assume valores inteiros. Por exemplo, o número de filhos ou o número de acidentes em uma estrada em um determinado dia.

Exemplo 1. *O volume utilizado de água em uma piscina é uma variável contínua. Imagine que a capacidade total de piscina é de 180 litros. Começando com a piscina vazia e adicionado água à piscina utilizando uma mangueira, o volume (em litros) utilizado após certo tempo pode assumir qualquer número real de 0 até 180 (inclusive valores não inteiros).*

Por outro lado, o número de pessoas que estão na piscina é uma variável discreta que pode assumir apenas valores inteiros não negativos.

Exemplo 2. *Um professor aplicou uma avaliação com 20 questões, na qual cada questão valia* 0,5 *pontos, de modo que a nota de cada aluno é um número no intervalo de* 0,0 *a* 10,0. *Mas veja que há apenas 21 notas possíveis:*

$$0{,}0, \quad 0{,}5, \quad 1{,}0, \quad \ldots \quad 8{,}5, \quad 9{,}0, \quad 9{,}5, \quad 10{,}0.$$

Apesar de que pode haver notas cujo valor não é inteiro, as notas não podem assumir qualquer valor no intervalo de 0 a 10. E os valores que elas podem assumir estão separados entre si (há saltos entre um número e outro). Desta forma, a nota obtida por um aluno nesta avaliação é uma variável discreta.

3 Frequência absoluta e relativa

A frequência absoluta e a frequência relativa são duas medidas importantes para descrever a *distribuição* de dados em um conjunto de observações. Elas descrevem, respectivamente, quantas vezes um determinado valor ocorre e qual é a proporção de ocorrências em relação ao total de observações.

Exemplo 3. *Foi feita uma pesquisa num grupo de 20 pessoas sobre o número de irmãos que cada tem. O entrevistador foi anotando cada uma das respostas, à medida que perguntava, produzindo a seguinte lista:*

$$0, 1, 2, 1, 4, 1, 2, 1, 0, 1, 0, 1, 2, 2, 4, 2, 1, 2, 4, 0.$$

A lista acima traz todas as informações obtidas na pesquisa, mas ela é confusa. Podemos organizar os dados em uma tabela para facilitar a análise. Primeiro, veja que apenas quatro respostas foram dadas: 0, 1, 2, *ou* 4 *(ninguém respondeu que tem 3 irmãos). Ademais:*

- *O número 0 aparece 4 vezes na lista;*
- *O número 1 aparece 7 vezes na lista;*
- *O número 2 aparece 6 vezes na lista;*
- *O número 4 aparece 3 vezes na lista.*

A quantidade de vezes que cada número aparece na lista é chamada de ***frequência absoluta*** *daquele número. Por exemplo, a frequência absoluta do número 0 é igual a 4, nesta pesquisa. A tabela abaixo mostra a frequência absoluta de cada resposta, de forma resumida.*

Número de irmãos	*Quantidade de observações*
0	*4*
1	*7*
2	*6*
4	*3*

Ao observar apenas esta tabela, não sabemos mais qual entrevistado respondeu qual número. Assim, temos uma versão resumida dos dados. Mas temos ideia muita mais clara do resultado geral da pesquisa. Por exemplo, podemos observar

que a resposta mais frequente é "1 irmão" (há 7 pessoas entrevistadas que possuem 1 irmão), seguida de "2 irmãos" (6 pessoas) por uma pequena diferença.

Para obter a frequência relativa *de cada resposta, dividimos a frequência absoluta pelo total de observações. No nosso exemplo, o total de observações é* $4 + 7 + 6 + 3 = 20$*, então a frequência relativa de cada resposta é:*

- *Frequência relativa de 0:* $\dfrac{4}{20} = 0{,}20 = 20\%$*;*
- *Frequência relativa de 1:* $\dfrac{7}{20} = 0{,}35 = 35\%$*;*
- *Frequência relativa de 2:* $\dfrac{6}{20} = 0{,}30 = 30\%$*;*
- *Frequência relativa de 4:* $\dfrac{3}{20} = 0{,}15 = 15\%$.

A tabela abaixo mostra a ***frequência relativa*** *de cada resposta, de forma resumida.*

Número de irmãos	*Proporção de observações*
0	20%
1	35%
2	30%
4	15%

Exemplo 4. *A tabela abaixo mostra a frequência absoluta cada um dos números obtidos ao lançar um dado de 6 faces, em um dado experimento. Calcule a tabela das frequências relativas de cada número.*

Número obtido	*Frequência absoluta*
1	*8*
2	*12*
3	*7*
4	*5*
5	*5*
6	*13*

Solução. A frequência relativa de cada número é obtida dividindo a frequência absoluta pelo total de observações, que é igual a $8+12+7+5+5+13=50$. Assim, temos a seguinte tabela:

Número obtido	Frequência relativa
1	$^{8}/_{50} = 16\%$
2	$^{12}/_{50} = 24\%$
3	$^{7}/_{50} = 14\%$
4	$^{5}/_{50} = 10\%$
5	$^{5}/_{50} = 10\%$
6	$^{13}/_{50} = 26\%$

□

Exemplo 5. *O conceito de frequência também pode ser aplicado a variáveis não quantitativas. Por exemplo, suponha que uma pesquisa foi feita com 40 pessoas, e cada uma delas foi questionada sobre sua cor preferida. As respostas foram as seguintes:*

- *8 pessoas responderam "azul";*
- *16 pessoas responderam "vermelho";*
- *4 pessoas responderam "verde";*
- *12 pessoas responderam "amarelo".*

A tabela abaixo mostra a frequência absoluta e a frequência relativa de cada resposta.

Cor preferida	*Frequência absoluta*	*Frequência relativa*
azul	*8*	20%
vermelho	*16*	40%
verde	*4*	10%
amarelo	*12*	30%

4 Média, moda e mediana

Média, moda e mediana são três conceitos extremamente importantes na estatística. Eles se aplicam apenas a *variáveis quantitativas*, sendo a "média" o mais utilizado entre eles. Aqui tratamos apenas do caso de variáveis quantitativas *discretas*.[1]

Eles são chamados de medidas de *tendência central*, pois são maneiras (diferentes) de encontrar um valor central que representa todo o conjunto de observações.

A **média** é a soma de todos os valores de um conjunto de dados dividida pelo número total de observações. É a medida mais comum de tendência central e é frequentemente chamada de "valor médio".

Exemplo 6. *Um grupo de 5 pessoas foi ao mercado e cada uma delas comprou certa quantidade de farinha. As quantidades compradas foram:* 2 kg, 5 kg, 1 kg, 3 kg *e* 4 kg. *A média da quantidade de farinha comprada por pessoa é:*

$$\frac{2+5+1+3+4}{5} = \frac{15}{5} = 3\,\text{kg}.$$

Neste exemplo, essa quantidade média é um número bastante importante para o dono da loja. Pois não interessa exatamente quanto cada cliente comprou, sabendo apenas a média e o número total de clientes por cada período, ele pode estimar a quantidade de farinha que deve comprar para manter sua loja abastecida. Por exemplo, se ele prevê que ao

[1] É possível falar da média de uma variável quantitativa contínua, mas isso requer ferramentas estudadas usualmente apenas no Ensino Superior (universitário).

longo da semana a loja irá receber 60 clientes interessados em farinha e que a média de consumo irá permanecer em $3\,\text{kg}$ *por cliente, ele deverá comprar* $60\times 3 = 180\,\text{kg}$ *de farinha (pois ao dividir o total de 180 quilos pelo 60 clientes, obtemos a média de* $180/60 = 3$ *quilos por cliente). Note que algumas pessoas podem comprar mais e outras menos, mas já temos uma boa estimativa de quanto será comprado ao todo.*

Observação 7. *Ao calcular uma média com o intuito de realizar uma estimativa, é importante que o dono da loja faça isso com um maior número de clientes e ao longo de vários horários, de vários dias da semana e mesmo em diferentes período do mês. Isso porque é possível que em um período específico do mês a quantidade de farinha comprada seja maior ou menor do que a média. Por exemplo, se a média for calculada apenas com os clientes de um dia de promoção, é possível que a média de vendas seja maior do que o normal. Ou se a média for calculada apenas com os clientes de um dia de chuva, é possível que a média seja menor do que o normal.*

Também é preciso ter uma boa estimava para o número de clientes que irão realizar a compra, para estimar o total de farinha necessária no estoque.

Exemplo 8. *Considere a seguinte lista representando as idades (em anos) de um grupo de 4 pessoas:* $15, 20, 30, 35$. *Temos que*

$$\frac{15+20+30+35}{4} = \frac{100}{4} = 25.$$

Portanto, a média das idades é 25 anos.

No exemplo acima, saber que a média das idades é de 25 anos nos dá alguma ideia das idades do grupo de pessoas, mas não tanto. Note que ninguém do grupo possui exatamente 25 anos (o que poderia acontecer, mas não acontece neste exemplo). Além disso há algumas pessoas bem mais novas e outras mais velhas. Os seguintes grupos de pessoas também possuem média de idade igual a 25, mas possuem características bastante diferentes: (i) um grupo de 6 pessoas

onde cada uma tem 25 anos de idade; (ii) um grupo onde há 3 pessoas com 5 anos de idade e 3 pessoas com 45 anos de idade. A média em cada um deles é:

$$\frac{25+25+25+25+25+25}{6} =$$

$$= \frac{5+5+5+45+45+45}{6} = \frac{150}{6} = 25.$$

Observe que a média é um número que sempre pertence ao intervalo que vai do menor ao maior número do conjunto de dados (podendo ser um dos extremos, caso todos o números sejam iguais). No entanto, a média nem sempre um bom indicador da tendência central. Por exemplo, considere o seguinte conjunto de idades: $\{20, 22, 25, 26, 27, 120\}$. A média das idades é:

$$\frac{20+22+25+26+27+120}{6} = \frac{240}{6} = 40 \text{ anos.}$$

No entanto, a maioria das pessoas tem idade entre 20 e 27 anos, e apenas uma pessoa tem idade de 120 anos (como 120 é muito maior do que os demais números isso acaba elevando a média).

Ao analisar o valor médio estamos fazendo uma análise bastante simplória. Afinal, estamos descrevendo uma lista de números por um único número, logo estamos perdendo informação. Isso não reduz a importância da média. Ela costuma ser o ponta pé inicial ideal da análise dos dados. Mas há outras medidas que também podemos adotar.

A **moda** é o valor que ocorre com mais frequência em um conjunto de dados. Em outras palavras, **é o valor que possui a maior frequência absoluta**. Um conjunto de dados pode ter uma única moda (unimodal) ou mais de uma moda (multimodal) se houver empate na frequência mais alta entre dois ou mais valores.

Exemplo 9. *Considere a seguinte lista representando as idades de um grupo de pessoas (em anos):*

$$25, 30, 25, 35, 30, 25, 45.$$

*Aqui, a idade "25" ocorre com mais frequência (3 vezes), enquanto as outras idades ocorrem apenas uma ou duas vezes. Portanto, este conjunto é unimodal e a **moda** desta lista é 25 anos.*

A **mediana** é o valor central de um lista de dados quando a mesma é ordenada em ordem crescente (ou decrescente). Ou seja, a mediana é o valor que divide o conjunto em duas partes iguais, (o mais próximo possível) de 50% dos valores abaixo dela e 50% dos valores acima dela. Para calcular a mediana, primeiro ordenamos os dados e, se houver um número ímpar de observações, a mediana será o valor do meio. Se houver um número par de observações, a mediana é calculada como a média dos dois valores do meio.

Observação 10. *Ao calcular o mediana, caso um número apareça várias vezes devemos listar todas as ocorrências deste número na lista.*

Exemplo 11. *Considere a mesma lista de idades do exemplo anterior:*

$$25, 30, 25, 35, 30, 25, 45.$$

Ordenando a lista obtemos:

$$25, 25, 25, 30, 30, 35, 45.$$

*O número que aparece no centro da lista é o número 30, logo a **mediana** é 30 anos.*

Exemplo 12 (Usando frequências absolutas para calcular a média)**.** *Considere mais uma vez a lista de números:*

$$25, 25, 25, 30, 30, 35, 45$$

Podemos calcular a média entre eles, diretamente, fazendo:

$$\frac{25+25+25+30+30+35+45}{7} = \frac{215}{7} = 30{,}71.$$

No entanto, podemos também calcular a média usando as frequências absolutas. Para isso, primeiro calculamos a frequência absoluta de cada número:

Número	*Frequência absoluta*
25	*3*
30	*2*
35	*1*
45	*1*

Para calcular a média iremos multiplicar cada número pela sua frequência absoluta, somar os resultados e, por fim, dividir pelo número total de observações (que coincide com a soma das frequências). Ou seja,

$$\frac{25 \cdot 3 + 30 \cdot 2 + 35 \cdot 1 + 45 \cdot 1}{3 + 2 + 1 + 1} = \frac{215}{7} = 30{,}71.$$

A expressão acima também é conhecida como média ponderada *dos números* $25, 30, 35, 45$ *tomando como pesos as respectivas frequências absolutas:* $3, 2, 1, 1$.

Observação 13. *No exemplo acima, veja que a soma*

$$25 \cdot 3 + 30 \cdot 2 + 35 \cdot 1 + 45 \cdot 1$$

é exatamente a mesmo que

$$\underbrace{25 + 25 + 25}_{3\ vezes} + \underbrace{30 + 30}_{2\ vezes} + 35 + 45.$$

Além disso, $3+2+1+1 = 7$*, de modo que a fração que representa a média é igual a* $215/7$*, nas duas formas de se calcular. Quando os números possuem frequências maiores, a expressão que envolve as frequências será muito mais compacta.*

Exemplo 14. *Certo jogo de tabuleiro teve seu preço reajustado várias vezes ao longo do ano, tanto devido à inflação como devido a promoções pontuais. A tabela abaixo traz quantas pessoas compraram o jogo para cada preço no qual houve alguma venda.*

Preço (em reais)	*Número de vendas*
100	30
110	10
120	20
130	10

Calcule o preço médio de venda do jogo.

Solução. Veja o total de vendas do jogo é igual a $30 + 10 + 20 + 10 = 70$ unidades. E o preço médio pode ser calculado usando a tabela de frequências absolutas:

$$\frac{100 \cdot 30 + 110 \cdot 10 + 120 \cdot 20 + 130 \cdot 10}{70} = \frac{7800}{70},$$

que é igual a aproximadamente 111,43 reais. □

Algo semelhante pode ser feito com frequências relativas.

Exemplo 15. *A tabela abaixo traz o percentual de pessoas que compraram certo telefone móvel para cada preço em que o mesmo foi vendido em uma loja. Calcule o preço médio das vendas.*

Preço (em reais)	*Percentual de vendas*
750	30%
900	30%
1000	40%

Solução 1. Podemos considerar que é um universo de 100 pessoas, nas quais 30 pagaram 750 reais, 30 pagaram 900 reais e 40 pagaram 1000 reais. Dessa forma a média do valor pago é de:

$$\frac{750 \cdot 30 + 900 \cdot 30 + 1000 \cdot 40}{100} =$$

$$= \frac{22.500 + 27.000 + 40.000}{100} = \frac{89500}{100} = 895 \text{ reais.}$$

□

Solução 2. Veja que $30\% + 30\% + 40\% = 0{,}3 + 0{,}3 + 0{,}4 = 1$. Assim, podemos calcular diretamente:

$$750 \cdot 0{,}3 + 900 \cdot 0{,}3 + 1000 \cdot 0{,}4 = 225 + 270 + 400 = 895 \text{ reais.}$$

□

Dicas para o Professor

Este material pode ser tratado em dois encontros de 50 minutos. Temos como pré-requisitos a familiaridade com número decimais, razões e proporções e a conversão de frações para porcentagens.